Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Maximum-likelihood density modification using pattern recognition of structural motifs

Thomas C. Terwilliger

Bioscience Division, Mail Stop M888, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Correspondence e-mail: terwilliger@lanl.gov

Received 17 April 2001
Accepted 17 August 2001

The likelihood-based approach to density modification [Terwilliger (2000), *Acta Cryst.* D**56**, 965–972] is extended to include the recognition of patterns of electron density. Once a region of electron density in a map is recognized as corresponding to a known structural element, the likelihood of the map is reformulated to include a term that reflects how closely the map agrees with the expected density for that structural element. This likelihood is combined with other aspects of the likelihood of the map, including the presence of a flat solvent region and the electron-density distribution in the protein region. This likelihood-based pattern-recognition approach was tested using the recognition of helical segments in a largely helical protein. The pattern-recognition method yields a substantial phase improvement over both conventional and likelihood-based solvent-flattening and histogram-matching methods. The method can potentially be used to recognize any common structural motif and incorporate prior knowledge about that motif into density modification.

1. Density modification by likelihood optimization

Although very powerful experimental methods exist for determining crystallographic phases in macromolecular crystallography, it is frequently necessary to improve or extend these phases before an atomic model of the macromolecule can be built. A variety of tools for density modification have been developed for this purpose, including solvent flattening, non-crystallographic symmetry averaging, histogram matching, phase extension, molecular replacement, entropy maximization and iterative model building (Abrahams & Leslie, 1996; Abrahams, 1997; Béran & Szöke, 1995; Bricogne, 1984, 1988; Cowtan & Main, 1993, 1996; Giacovazzo & Siliqi, 1997; Goldstein & Zhang, 1998; Gu *et al.*, 1997; Lunin, 1993; Perrakis *et al.*, 1997; Podjarny *et al.*, 1987; Prince *et al.*, 1988; Refaat *et al.*, 1996; Roberts & Brünger, 1995; Rossmann & Arnold, 1993; Shneerson *et al.*, 2001; Szöke, 1993; Szöke *et al.*, 1997; Terwilliger, 2000; Vellieux *et al.*, 1995; Wilson & Agard, 1993; Xiang *et al.*, 1993; Zhang & Main, 1990; Zhang, 1993; Zhang *et al.*, 1997). The basis of density modification is that there are many possible sets of phases that are reasonably consistent with the experimental data and the most likely of these sets of phases are those that lead to electron-density maps that are most consistent with expectations for a macromolecule. The most common way to carry out density modification has been to calculate an electron-density map, modify it to meet expectations, calculate modified phases and combine the modified phases with experimental phases to yield new estimates of the crystallographic phases. This method has the disadvantages that optimal weighting of

modified and experimental phases is difficult and that it is not clear when to stop iterating. The difficulty in weighting in particular is well known and a number of approaches have been designed to circumvent it, including the use of maximum-entropy methods and the use of weighting optimized using cross-validation (Xiang *et al.*, 1993; Roberts & Brünger, 1995; Cowtan & Main, 1996) and 'solvent flipping' (Abrahams & Leslie, 1996).

We have recently developed a method for carrying out density modification that consists of directly maximizing the likelihood of the structure factors, including both experimental information and the characteristics of the electron density resulting from the structure factors (Terwilliger, 1999, 2000). The general idea is very simple. We express the total likelihood of a set of structure factors $\{F_{\mathbf{h}}\}$ in terms of three quantities: (i) any prior knowledge we have from other sources about these structure factors, (ii) the likelihood that we would have measured the observed set of structure factors $\{F_{\mathbf{h}}^{\rm OBS}\}$ if this set of structure factors were correct and (iii) the likelihood that the map resulting from this set of structure factors $\{F_{\mathbf{h}}\}$ is consistent with our prior knowledge about this and other macromolecular structures. This can be written as

$$\mathrm{LL}(\{F_{\mathbf{h}}\}) = \mathrm{LL}^{o}(\{F_{\mathbf{h}}\}) + \mathrm{LL}^{\rm OBS}(\{F_{\mathbf{h}}\}) + \mathrm{LL}^{\rm MAP}(\{F_{\mathbf{h}}\}), \quad (1)$$

where $\mathrm{LL}(\{F_{\mathbf{h}}\})$ is the log-likelihood of a possible set of crystallographic structure factors $F_{\mathbf{h}}$, $\mathrm{LL}^{o}(\{F_{\mathbf{h}}\})$ is the log-likelihood of these structure factors based on any information that is known in advance, such as the distribution of intensities of structure factors (Wilson, 1949), $\mathrm{LL}^{\rm OBS}(\{F_{\mathbf{h}}\})$ is the log-likelihood of these phases given the experimental data alone and $\mathrm{LL}^{\rm MAP}(\{F_{\mathbf{h}}\})$ is the log-likelihood of the electron-density map resulting from these phases. In this formulation, density modification consists of maximizing the total likelihood given by (1). To maximize this likelihood, it is necessary both to define a map-likelihood function and to have a practical way of finding structure factors that maximize it.

We recently developed a formulation of the map-likelihood function that often allows a straightforward and rapid optimization of the total likelihood in (1). The log-likelihood for the electron-density map $\mathrm{LL}^{\rm MAP}(\{F_{\mathbf{h}}\})$ is written as the integral over the map of a local log-likelihood of electron density, $\mathrm{LL}[\rho(\mathbf{x}, \{F_{\mathbf{h}}\})]$,

$$\mathrm{LL}^{\rm MAP}(\{F_{\mathbf{h}}\}) \simeq \frac{N_{\rm REF}}{V} \int_{V} \mathrm{LL}[\rho(\mathbf{x}, \{F_{\mathbf{h}}\})]\, \mathrm{d}^3\mathbf{x}. \quad (2)$$

This formulation neglects contributions to the log-likelihood of the map that involve more than one point at a time, but is nevertheless very useful in describing the overall likelihood of the map (Terwilliger, 1999, 2000).

As long as the first and second derivatives of the local log-likelihood of electron density with respect to electron density can be calculated, a steepest-ascent method can be used to optimize the total likelihood in (1) (Terwilliger, 1999, 2000). In this broad class of situations, an FFT-based method can be used to approximate derivatives of the total map log-likelihood function with respect to each structure factor (Terwilliger, 1999, 2000). These derivatives can then in turn be used in a Taylor's series expansion to approximate the total map log-likelihood function as a function of each structure factor. This makes it practical to optimize the total likelihood in (1) because the other terms (*a priori* knowledge of phases, and experimental phase information) are also normally expressed separately for each structure factor. In each cycle of optimization, a new probability distribution for each structure factor (or phase) is obtained by calculating the relative likelihood of each possible value of that structure factor using (1) with this approximation for the map log-likelihood function.

The local map log-likelihood function in (2) is a critical element in our maximum-likelihood density-modification approach. This likelihoood function could include any type of expectations about the electron-density value at a particular point in the map. In particular, we have shown that expectations about electron-density values at points both in the solvent region and in the protein region of a protein crystal can be included in maximum-likelihood density modification and that this approach can be very powerful for improving crystallographic phases (Terwilliger, 1999, 2000). We show here that the same approach can be used to incorporate detailed information about patterns of electron density in a map such as those corresponding to secondary-structural elements in a protein structure.

2. Local log-likelihood function for a map

The local map log-likelihood function is essentially a statement of the plausibility of each possible value of electron density at a point in the electron-density map. It is important to recognize that for the present purpose this probability of electron density is in the context of all the errors in the map caused by uncertainty in structure factors (Terwilliger, 2000). This distinction is necessary because in any one cycle of our approach each phase is optimized independently of all others. Consequently, as one phase (or structure factor) is being optimized it is in the context of the errors remaining in all other phases. This means that even in an idealized case in which the value of the true electron density was known exactly at a particular point in the map, the correct value of a particular phase would not ordinarily lead to exactly this value of electron density. Instead, the probability distribution of plausible electron densities at this point would have a finite breadth corresponding to the overall error in the map.

Following this approach, the probability distribution $p(\rho)$ for electron density at the point $\mathbf{x}$ in a map with substantial phase errors can be written as

$$p(\rho) = \int_{\rho_T} p(\rho_T) \exp\left[-\frac{(\rho - \beta\rho_T)^2}{2\sigma^2_{\rm MAP}}\right] \mathrm{d}\rho_T, \quad (3)$$

where $p(\rho_T)$ is the probability distribution for electron density in a model (perfect) case, $\sigma^2_{\rm MAP}$ is the variance in the map and β is a scale factor (Terwilliger, 2000).

As it is generally not known for certain whether a particular point $\mathbf{x}$ is in the protein or solvent region, it is useful to write the local map-likelihood function as the sum of conditional

probabilities dependent on which environment the point is located in,

$$\mathrm{LL}[\rho(\mathbf{x}, \{F_{\mathbf{h}}\})] = \ln\{p[\rho(\mathbf{x})|\mathrm{PROT}]p_{\mathrm{PROT}}(\mathbf{x}) + p[\rho(\mathbf{x})|\mathrm{SOLV}]p_{\mathrm{SOLV}}(\mathbf{x})\}, \quad (4)$$

where $p_{\mathrm{PROT}}(\mathbf{x})$ is the probability that $\mathbf{x}$ is in the protein region, $p[\rho(\mathbf{x})|\mathrm{PROT}]$ is the conditional probability for $\rho(\mathbf{x})$ given that $\mathbf{x}$ is in the protein region and $p_{\mathrm{SOLV}}(\mathbf{x})$ and $p[\rho(\mathbf{x})|\mathrm{SOLV}]$ are the corresponding quantities for the solvent region. The probability that $\mathbf{x}$ is in the protein or solvent regions can estimated by a modification of the methods of Wang (1985) and Leslie (1987) as described earlier (Terwilliger, 1999).

3. Incorporating information obtained from image reconstruction

The local log-likelihood function for the map in (4) is based simply on probability distributions for the protein and solvent regions of the map. The same approach can be applied to information on the likely values of electron density at a particular point derived from any other source. In particular, suppose that it were known that there is the probability p_H that there is a helix in a particular orientation located at a particular place in the unit cell. Then our prior knowledge about the electron-density distribution in a helix could be used in just the same way as our knowledge about the electron density in the solvent region of the unit cell. At each point within and in the immediate vicinity of this helix, a probability distribution for plausible values of electron density could be constructed using model values of electron density for a helix along with (3). These probability distributions could then be used in a local log-likelihood function that is an extension of (4):

$$\mathrm{LL}[\rho(\mathbf{x}, \{F_{\mathbf{h}}\})] = \ln\{p[\rho(\mathbf{x})|\mathrm{PROT}]p_{\mathrm{PROT}}(\mathbf{x}) + p[\rho(\mathbf{x})|\mathrm{SOLV}]p_{\mathrm{SOLV}}(\mathbf{x}) + p[\rho(\mathbf{x})|H]p_H(\mathbf{x})\}, \quad (5)$$

where $p_H(\mathbf{x})$ refers to the probability that there is a helix at a known location, with a known orientation, somewhere near the point $\mathbf{x}$; $p[\rho(\mathbf{x})|H]$ is the probability distribution for electron density at this point given that this helix actually is present. As there is nothing special about helices (other than their relative regularity), (5) could equally well be used to include any other type of structural motif or indeed any other pattern of electron density that can be recognized. The significance of (5) is that it provides a way to incorporate pattern recognition (the probability that there is a helix with this orientation at this point) into density modification. If the pattern to be detected involves a large part of the map, then it might be identifiable even when errors in the map are very large. Then if the pattern is well defined the last term in (5) could potentially contribute very substantially to the local log-likelihood function and therefore to density modification. This approach can be thought of as a likelihood-based extension of the iterative skeletonization procedure for phase improvement (Baker *et al.*, 1993; Wilson & Agard, 1993) and of the iterative model-building procedures incorporated into *ARP* and *wARP* (Perrakis *et al.*, 1999).

The formulation in (5) essentially segments the map into points within protein, within solvent and within another pattern (helix) of electron density. Strictly speaking, these categories are clearly not mutually exclusive, as a point can be both within protein and within a helix. Furthermore, a particular point could be within more than one helix pattern (as the template used to identify a helix might be shorter than the actual helix and several overlapping patterns of helix might be recognized). It is convenient, however, to use the most informative piece of information when there is either type of overlap. If a point is both within the protein region and within a helix, for example, the fact that it is within a helix is far more informative because it defines the electron density very precisely, while the fact that the point is within the protein only gives a very broad range for possible values of electron density. In practice, if more than one pattern has information about the electron density at a particular point, then the pattern that has the highest probability is used. Then the probabilities that the point is in protein or solvent are modified from our earlier expressions (Terwilliger, 1999, 2000) by normalizing their total to simply be whatever the probability is that the point is not in this pattern.

4. Image reconstruction by template matching

Template matching has been used as an aid to map interpretation for some time in X-ray crystallography (Kleywegt & Jones, 1997, 1998; Cowtan, 1998). Many structural elements in proteins are quite uniform and can sometimes be recognized in even a noisy electron-density map. In the context of image reconstruction, once an element such as a helical region is recognized, the electron density in the neighborhood of the main-chain atoms can often be estimated more accurately from the model of a helix than from the map itself.

To make optimal use of (5), a method is needed for estimating the probability that a particular pattern of electron density (*e.g.* one corresponding to a helix) is located at each possible position and with each possible orientation in the unit cell. To make this practical, it is convenient to separate it into three steps. First, a template is constructed that is an average of the patterns of electron density found in many instances where it occurs. Next, locations and orientations of a template (such as the electron density for a helix) that match the electron density in the map to some degree are identified. Then the probabilities of these possibilities are estimated.

4.1. Construction of a template for a helix

Although helices are relatively regular secondary structures, there is some variation from one to another in the precise locations of atoms and in their thermal factors. Even more importantly, the side chains in one helix may be completely different to those in another. Consequently, construction of a template that has average features is useful for the purpose of pattern matching. Additionally, it is helpful

to have a point-by-point estimate of the standard deviation of this density that can be used to identify regions within the template that have more or less variation. We used a simple method to generate a template and standard deviation of the template for helices. Residues 133–138 of myoglobin (PDB entry 1a6m) were chosen as a model helical segment. Then 326 segments of six amino acids from the largely helical protein phycoerythrin (Chang *et al.*, 1996; PDB code1lia) for which the N, C, C^{α} and O atoms could be superimposed on the corresponding atoms in the myoglobin helix with an r.m.s. deviation of 0.5 Å or less were used to generate an average template for α-helices.

The template was constructed by superimposing each six-amino-acid helical segment of phycoerythrin on the myoglobin helix and calculating an electron-density map at a resolution of 3 Å based on all atoms of the phycoerythrin structure that fell inside a 20 Å cube with the helix at the center. The resulting electron density within 2.5 Å of an atom in the myoglobin helix was averaged to yield our helical template. The average density in the template region was adjusted to a value of zero and all points outside the template region were set to values of zero. At the same time, the standard deviation of electron density at the same set of points was determined.

Fig. 1 shows the resulting helical template. The positions of C^{β} atoms are visible, but all further side-chain atoms are sufficiently different at different positions that no density is visible.

4.2. Matching a helix template to an electron-density map

We used an FFT-based convolution method to identify rotations and translations of our helix template that match the electron density in a map to some degree (Kleywegt & Jones, 1997; Cowtan, 1998). The helix template was rotated in real space and placed at the origin of a unit cell with dimensions identical to the map to be searched. Structure factors for the rotated template were calculated in space group $P1$ and the convolution of the template and the electron-density map was calculated using an FFT. Each point in this convolution corresponds to a translation of the rotated template. The value of the convolution at each point is essentially the integral over the template region of the density in the rotated translated template, multiplied by the density in the map. This product is expected to be high if the rotated translated template has a high correspondence to the map and low otherwise.

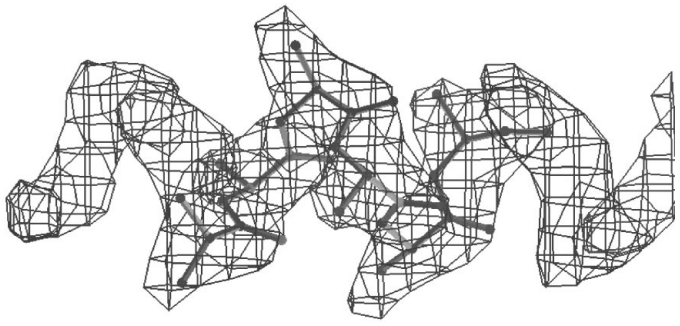

Figure 1
Averaged helical template. The template was calculated at a resolution of 3 Å as described in the text.

In our implementation of a helix search, the template is rotated in increments of 10° over three rotation axes. As our α-helix template is essentially symmetric when rotated 100° about its axis, the search only included 100° of rotation about the helix axis.

To identify peaks in this search that are reasonably likely to correspond to actual helical segments in the electron-density map, a height cutoff was calculated such that in a random map only about one peak would be chosen every other rotation. The cutoff was estimated from the number of reflections (an estimate of the number of degrees of freedom in the map), the mean and standard deviation of the convolution function. Typically, the cutoff was in the range of 3σ to 4σ and typically about 200–2000 peaks were saved. In cases where there are templates with center-to-center distances of less than 2 Å, the one with the higher peak height was chosen.

Once matches of template to map are identified in this fashion, the rotation and translation parameters are refined by minimizing the residual error in the fit between the map and the template. This residual error $\sigma_{\rm RESID}$ is estimated from the r.m.s. difference $\sigma_{\rm FIT}$ between the map and the template (after multiplying the template by a scale factor α and adding an adjustable offset) and the uncertainty in the template itself σ_H (based on the variability in electron densities for model helices),

$$\sigma^2_{\rm RESID} = \sigma^2_{\rm FIT} - (\alpha\sigma_H)^2. \qquad (6)$$

4.3. Estimating probabilities of matches of a template to a map

In the scheme described above (5) for incorporating information about patterns of electron density in a map, it is essential to have an estimate of the probability p_H that a template is actually located at a particular position and with a particular orientation. The convolution-based search we use to identify plausible matches is not entirely suitable for this purpose because the peak heights are just a measure of how good the match is, not how likely it is that this pattern really is located there. To see the difference, consider a case where it is known somehow that there are no helices of six amino acids in length in a particular protein, but where there is a stretch of three amino acids in an α-helical conformation. A convolution-based search might show a large peak corresponding to overlap of the template and these three amino acids, yet only part of the template pattern is really present. In this example, it might be reasonable to say that there is a 50% chance that any given point in the template is a good description of the true electron density in the map, but not to say that this chance is 100%.

We use a combination of prior knowledge of the helical content of the protein in the crystal and the correlation coefficient of each match of template to map to estimate the

probability that each match correctly identifies a region of the map with this pattern of electron density. First, the mean $\overline{\mathrm{CC}}$ and standard deviation σ_{CC} of correlation coefficients were determined for randomly chosen template orientations and translations. This allows an estimate for each match of template to map of the probability $p(\mathrm{CC_{OBS}}|\mathrm{not}\ H)$ that this match with a correlation coefficient of $\mathrm{CC_{OBS}}$ would have occurred entirely by chance (that is, if there were no helical pattern at this location),

$$p(CC_{\mathrm{OBS}}|\mathrm{not}\ H) \propto \exp\left[-\frac{(\mathrm{CC_{OBS}} - \overline{\mathrm{CC}})^2}{2\sigma_{\mathrm{CC}}^2}\right]. \quad (7)$$

Next, we estimate the number of templates that are likely to be needed to describe all the helical regions in the unit cell. This is necessarily rather approximate both because the number of residues in helical conformation is not ordinarily known very accurately and because in our method the templates describing a helix can overlap. Using the prior knowledge of the fraction f_H of the macromolecule that is in a helical conformation and of the fraction f_{PROT} of the unit cell that is occupied by macromolecule, the cell volume V and the template volume V_{template}, and using the empirical observations that about 70% (f_{template}) of the volume in a model helical protein is within a corresponding helical template and that only about 35% (f_{unique}) of each template does not overlap with another template, we can write that

$$N_{\mathrm{template}} \simeq \frac{f_H f_{\mathrm{PROT}} f_{\mathrm{template}} V}{f_{\mathrm{unique}} V_{\mathrm{template}}}. \quad (8)$$

Now we can estimate the relative probability $p(H|\mathrm{CC_{OBS}})$ that each template match, with correlation coefficient $\mathrm{CC_{OBS}}$, is at least partially correct (that is, it does not arise by chance),

$$p(H|\mathrm{CC_{OBS}}) = \frac{p_o(H)\ p(\mathrm{CC_{OBS}}|H)}{p_o(H)\ p(\mathrm{CC_{OBS}}|H) + p_o(\mathrm{not}\ H)\ p(\mathrm{CC_{OBS}}|\mathrm{not}\ H)}, \quad (9)$$

where $p_o(H)$ and $p_o(\mathrm{not}\ H)$ are the *a priori* probabilities that there is or is not a helix located at this position and orientation and $p(\mathrm{CC_{OBS}}|H)$ and $p(\mathrm{CC_{OBS}}|\mathrm{not}\ H)$ are the probabilities that this correlation coefficient would be found for correct and incorrect matches, respectively. As the vast majority of locations and orientations do not correspond to a correct match, we can reasonably assume that $p_o(\mathrm{not}\ H) \simeq 1$. Additionally, as we are only considering the highest peaks in the convolution, it is reasonable to assume that correct matches could have led to any of the peak heights observed, so that $p(\mathrm{CC_{OBS}}|H) \simeq 1$. As we have an expression for $p(\mathrm{CC_{OBS}}|\mathrm{not}\ H)$ (7), the only unknown term in (9) is $p_o(H)$, the *a priori* probability that there is a helix in this position and orientation. We estimate $p_o(H)$ by adjusting it so that the total number of templates is equal to N_{template} (7–9):

$$N_{\mathrm{template}} = \sum_{\mathrm{templates}} p(H|\mathrm{CC_{OBS}}), \quad (10)$$

where the probability that each template match is at least partially correct is

$$p(H|\mathrm{CC_{OBS}}) = \frac{p_o(H)}{p_o(H) + \exp-[(\mathrm{CC_{OBS}} - \overline{\mathrm{CC}})^2/(2\sigma_{\mathrm{CC}}^2)].} \quad (11)$$

Although all possible matches with all levels of probability might ideally be included in the image-reconstruction process, we find that in practice only the most probable ones contribute in a useful way. Consequently, only template matches with a value of $p(H|\mathrm{CC_{OBS}}) > 0.8$ are included.

Finally, as discussed above there may be many cases where part of the template matches a pattern in the map but another part does not. We estimate this fraction that matches the pattern (f_{match}) based on the ratio of the correlation coefficient for each match ($\mathrm{CC_{OBS}}$) to the highest correlation coefficient for any match in the map ($\mathrm{CC_{MAX}}$),

$$f_{\mathrm{match}} \simeq \frac{\mathrm{CC_{OBS}}}{\mathrm{CC_{MAX}}}. \quad (12)$$

Using (11) along with the average helix template and its standard deviation, we are now in a position to evaluate the new terms in (5). The probability $p_H(\mathbf{x})$ that there is a helix at a particular location and orientation that contributes some information about the electron density at point $\mathbf{x}$ is given by

$$p_H(\mathbf{x}) \simeq f_{\mathrm{match}} p(H|\mathrm{CC_{OBS}}), \quad (13)$$

where the probability that this template match is at least partially correct is $p(H|\mathrm{CC_{OBS}})$ (11), where the estimated fraction of the template that is involved in the match is f_{match} and where H refers to a template match that overlaps the point $\mathbf{x}$. The probability distribution for electron density at $\mathbf{x}$ is given by (3), where the ideal electron-density distribution $p(\rho_T)$ is based on the mean ρ_{template} and standard deviation $\sigma_{\mathrm{template}}$ of the rotated translated template at the point $\mathbf{x}$,

$$p(\rho_T) \simeq \exp -\left[\frac{(\rho_T - \beta\rho_{\mathrm{template}})^2}{2\sigma_{\mathrm{template}}^2}\right]. \quad (14)$$

5. Application to density modification of a map of an α-helical protein

We tested our pattern-matching approach to density modification using the armadillo repeat region of β-catenin, which is largely α-helical (Huber *et al.*, 1997). This structure was solved using MAD phasing on 15 Se atoms incorporated into methionine residues in the protein. To make the test suitably difficult, we used only three of the 15 Se atoms in calculating initial phases. As expected, this led to a very noisy map; the correlation coefficient of this map with a map calculated using phases from the refined model was only 0.29 (Fig. 2*a*). We carried out real-space density modification using *DM* (Cowtan & Main, 1996), resulting in some improvement of the map and a correlation coefficient of only 0.42 (not shown). The maximum-likelihood density-modification approach we described earlier (without any pattern recognition) resulted in a substantial improvement in the map, with a correlation coefficient of 0.62 (Fig. 2*b*). The pattern recognition of helices is illustrated in Fig. 2(*c*). In order to visualize the templates,

the map shows the electron density in the rotated translated templates (from Fig. 1), multiplied by the probability that the template is a correct match (13). The density in the templates is a fairly good but not perfect match to the refined atomic model. The maximum-likelihood density modification with pattern recognition of helices improved the map even more substantially, with an overall correlation coefficient of 0.67 (Fig. 2*d*).

An even more difficult map to interpret is illustrated in Fig. 3. This map was created in the same way as the one in Fig. 2, except that only one selenium was used in phasing the 700 amino-acid residue protein. The starting correlation coefficient of the map with the model map was just 0.24; maximum-likelihood density modification increased this to 0.32 and density modification with pattern recognition to 0.51.

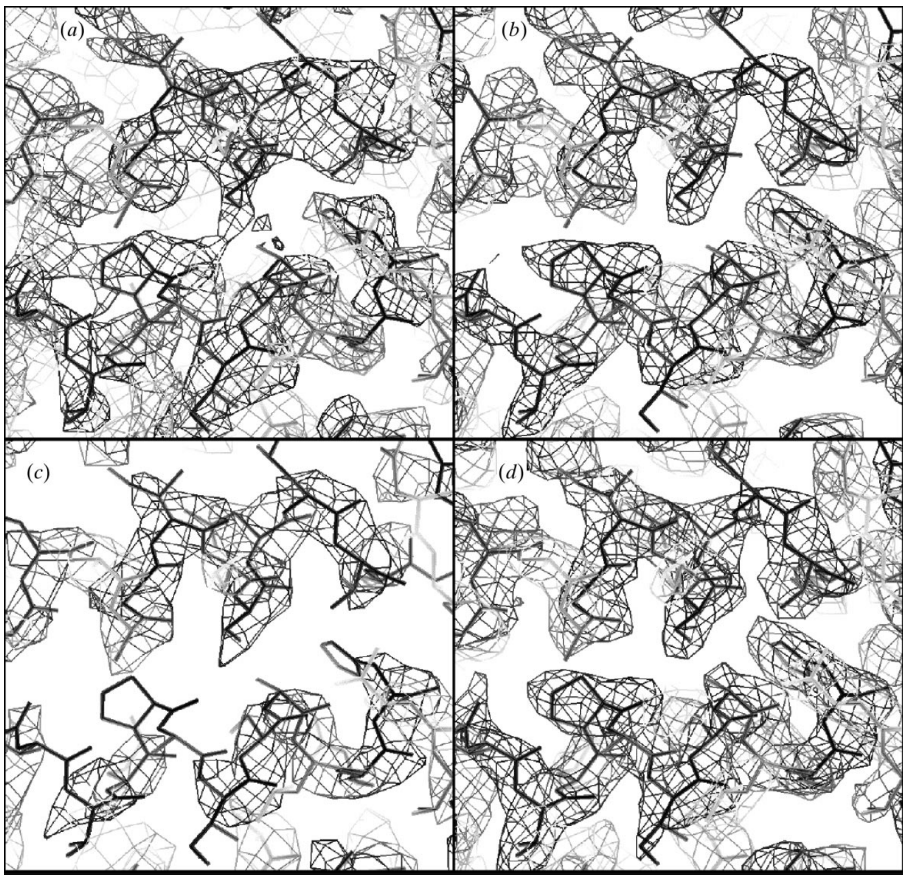

Figure 2
Experimental, real-space density-modified, maximum-likelihood density-modified and maximum-likelihood with pattern-recognition modified maps of an α-helical protein. The armadillo repeat region of β-catenin crystallizes in space group $C222_1$, with unit-cell parameters $a = 64$, $b = 102$, $c = 187$ Å and a solvent content of about 50% (Huber *et al.*, 1997). Phases were calculated with *SOLVE* (Terwilliger & Berendzen, 1999) using three selenium sites at a resolution of 3 Å. A section of this map is shown in (*a*). Real-space density modification was carried out with *DM* (Cowtan & Main, 1996) using solvent flattening with a solvent content of 50% and histogram matching (not shown). Maximum-likelihood density modification without image reconstruction was carried out as described earlier (Terwilliger, 2000) using a solvent content of 50% (*b*). Templates found in the experimental electron-density map are illustrated in (*c*). Maximum-likelihood density modification with pattern recognition was carried out as described in the text, using a solvent content of 50% and a fraction helical secondary structure of 80% (*d*). Template matches with a probability less than 0.8 were not included.

6. Discussion

The density-modification procedures developed here and in our recent work (Terwilliger, 1999, 2000) contain two substantial changes from existing methods. One is the use of optimization of a likelihood function rather than phase recombination between experimental and modified maps. The second is the use of a log-likelihood function for a map.

The optimization of a likelihood function (more precisely a posterior probability function in this case, *e.g.* equation 1) is important, as discussed in depth by others (Bricogne, 1984, 1988; Lunin, 1993), because it places density modification on a sound statistical foundation. In the present case, it also eliminates difficulties in weighting of experimental and modified phases. This optimization is made practical by the approaches we have developed involving reciprocal-space calculations of derivatives of the likelihood function with respect to structure factors.

A more far-reaching change from existing methods is in the development of a likelihood function for a map. This likelihood function is a statement of the plausibility of an electron-density map calculated from some set of structure factors. The plausibility can include any information about patterns of electron density that are expected and not expected. Our implementation of the likelihood function for a map (2) is a simplified version in which each point in the map is treated independently. The overall log-likelihood of the map is the integral over the unit cell of the local map log-likelihood function.

The use of a map-likelihood function is related to the methods of Szöke (1993; Szöke *et al.*, 1997) and Béran & Szöke (1995) in which crystallographic phases are obtained by matching the electron density in a part of the unit cell to a target value. The maximum-likelihood approach described here differs from these methods in that probabilistic descriptions of the expected electron density are used, allowing a calculation of phase probability distributions, rather than searching for a set of phases that is consistent with constraints.

The local log-likelihood function for a map can readily incorporate information about solvent and protein regions in the map if they are identified by some means (Terwilliger, 2000). After taking into consideration the noise in the map (3), the electron density at a point known to be in the solvent region is plausible only

if it has values within a narrow range expected in the solvent. Similary, the density at a point in the protein region is plausible only if it has a value in the somewhat greater range expected in the protein region.

The patterns of electron density that are included in the local log-likelihood function need not be as simple as the probability distribution for electron density in solvent or protein regions. They can also include detailed information about the electron density in a region. (5) shows how to incorporate information on a pattern of density corresponding to a structural motif such as a fragment of α-helix. Any other fragment density information could be incorporated in a similar fashion.

It is important to recognize that the use of partial structure information in a likelihood function for a map is fundamentally different than using what may appear to be the same partial structure information in a σ_A or related model phase calculation (Read, 1986). The difference is that in the σ_A model phase calculation, the errors in the partial structure information are assumed to be the same everywhere in the unit cell, while in the map-likelihood approach, the errors can be explicitly specified for each point in the map. The method of Szöke (1993) also has this property.

The difference can be best appreciated in an idealized case where a only small fragment of structure is missing from an otherwise perfect model and a difference Fourier or similar calculation is carried out to identify the missing fragment. In the σ_A-weighted map, the difference density can be located anywhere in the map (though much will be in the correct region). In the map-likelihood approach, the fact that the density is known exactly everywhere except in the region of the missing fragment is explicitly taken into account. Consequently, in this approach all the difference density would be located in the region where the missing fragment is located.

In a more accessible case the same principle applies as well. In the examples described in this work, α-helices are identified in a map and used to improve phases. In the model phase calculation approach, the rotated translated templates (or coordinates of atoms in a model helix) would be used to calculate model phases and a σ_A-weighted combined phase map would be calculated. As in the more extreme example above, the uncertainties in electron density based on the model alone would be assumed to be distributed over the entire unit cell. In the map-likelihood method, uncertainties in electron density are relatively low in the entire region of each helical template (where the model electron density is relatively well known) and higher elsewhere in the protein region (where it is poorly known) and once again lower in the solvent region (where it is very precisely known). This point-by-point specification of uncertainty in the map allows a much more complete use of the available information about the partial model than the model phase method.

The key to the use of the local log-likelihood function for a map is the specification of a probability distribution for the electron density for some subset of points in the map. It does not matter if this specification says that all the points in a region have the same electron density or whether the points in this region have a particular pattern of electron density such as a part of a helix. Much the same amount of information is conveyed in either case and essentially the same amount of improvement in phases or structure factors can potentially be obtained in either case.

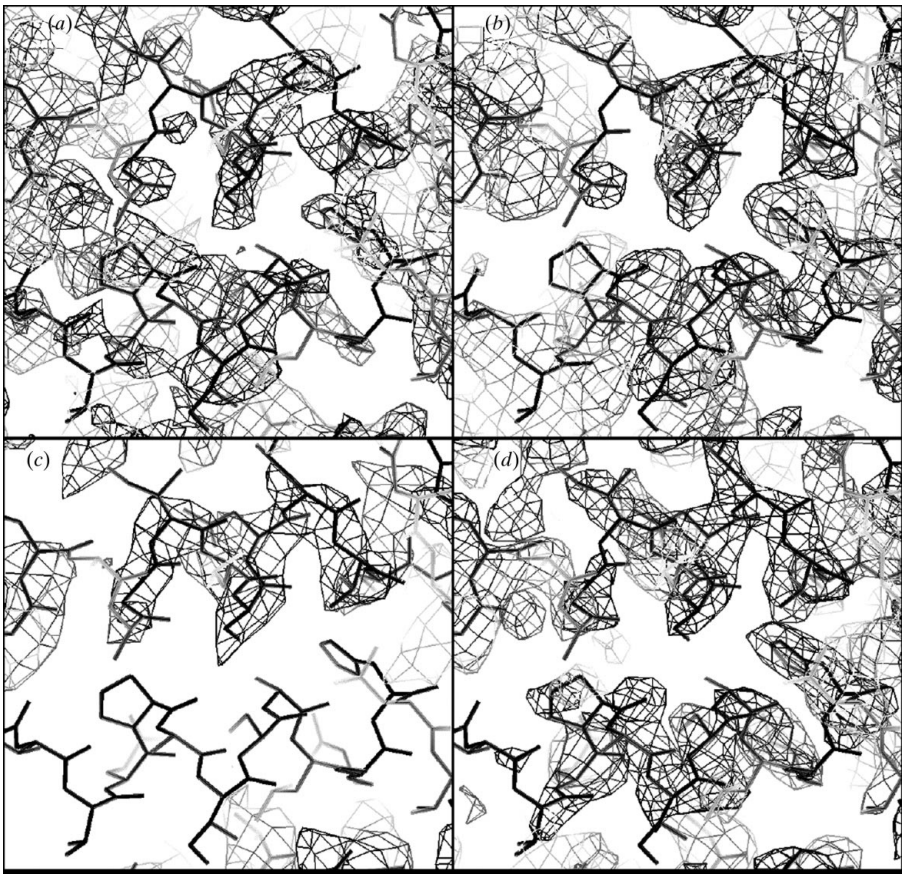

Figure 3
Template matching with a very noisy map. Analyses were carried out as in Fig. 2, starting with a map calculated using one selenium for phasing.

7. Conclusions

The methods we have developed here and in recent work (Terwilliger, 1999, 2000) provide a simple and practical way to incorporate prior knowledge of the electron density in a crystal structure into probability distributions for structure factors. The prior knowledge can range from the locations of solvent and protein regions to detailed infor-

mation on a local pattern of electron density corresponding to a fragment of structure.

There are a number of important extensions of our approaches that can be readily envisioned. One is the incorporation of non-crystallographic symmetry information. Electron-density information from one copy of a macromolecule in the asymmetric unit can be used in our approach in the same way as other partial structure information. The ability to specify separate probability distributions for electron density at each point in the map will make it possible to take into account the different amounts of error in different parts of the partial model. In that way, the parts that are most similar can effectively be weighted more strongly and the parts that are more different be weighted less strongly, a property that is more difficult to achieve with current methods.

A second is in the area of molecular replacement. The calculation of phases from a partial model is currently problematic owing to model bias. The ability to specify, on a point-by-point basis, the uncertainties in a model could substantially improve the quality of phasing that can be obtained. A third is in automated model building. The approach described here for identification of α-helices and incorporation of model information into density modification is essentially the first step in automated model building. The iterative approaches incorporated into *ARP* and *wARP* (Perrakis *et al.*, 1999) could be modified to incorporate the likelihood functions we have described here.

The author would like to thank Joel Berendzen for helpful discussions and the NIH and the US Department of Energy for generous support.

References

Abrahams, J. P. (1997). *Acta Cryst.* D**53**, 371–376.
Abrahams, J. P. & Leslie, A. G. W. (1996). *Acta Cryst.* D**52**, 30–32.
Baker, D., Bystroff, C., Fletterick, R. J. & Agard, D. A. (1993). *Acta Cryst.* D**49**, 429–439.
Béran, P. & Szöke, A. (1995). *Acta Cryst.* A**51**, 20–27.
Bricogne, G. (1984). *Acta Cryst.* A**40**, 410–445.
Bricogne, G. (1988). *Acta Cryst.* A**44**, 517–545.
Chang, W. R., Jiang, T., Wan, Z. L., Zhang, J. P., Yang, Z. X. & Liang, D. C. (1996). *J. Mol. Biol.* **262**, 721–731.
Cowtan, K. D. (1998). *Acta Cryst.* D**54**, 750–756.
Cowtan, K. D. & Main, P. (1993). *Acta Cryst.* D**49**, 148–157.
Cowtan, K. D. & Main, P. (1996). *Acta Cryst.* D**52**, 43–48.
Giacovazzo, C. & Siliqi, D. (1997). *Acta Cryst.* A**53**, 789–798.
Goldstein, A. & Zhang, K. Y. J. (1998). *Acta Cryst.* D**54**, 1230–1244.
Gu, Y., Zheng, C., Zhao, Y., Ke, H. & Fan, H. (1997). *Acta Cryst.* D**53**, 792–794.
Huber, A. H., Nelson, W. J. & Weis, W. I. (1997). *Cell*, **90**, 871–882.
Kleywegt, G. J. & Jones, T.A. (1997). *Acta Cryst.* D**53**, 179–185.
Kleywegt, G. J. & Jones, T.A. (1998). *Acta Cryst.* D**54**, 1119–1131.
Leslie, A. G. W. (1987). *Proceedings of the CCP4 Study Weekend*, pp. 25–31. Warrington: Daresbury Laboratory.
Lunin, V. Y. (1993). *Acta Cryst.* D**49**, 90–99.
Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
Perrakis, A., Sixma, T. K., Wilson, K. S. & Lamzin, V. S. (1997). *Acta Cryst.* D**53**, 448–455.
Podjarny, A. D., Bhat, T. N. & Zwick, M. (1987). *Annu. Rev. Biophys. Biophys. Chem.* **16**, 351–373.
Prince, E., Sjolin, L. & Alenljung, R. (1988). *Acta Cryst.* A**44**, 216–222.
Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.
Refaat, L. S., Tate, C. & Woolfson, M. M. (1996). *Acta Cryst.* D**52**, 252–256.
Roberts, A. L. U. & Brünger, A. T. (1995). *Acta Cryst.* D**51**, 990–1002.
Rossmann, M. G. & Arnold, E. (1993). *International Tables for Crystallography*, Vol. B, edited by U. Shmueli, pp. 230–258. Dordrecht: Kluwer Academic Publishers.
Shneerson, V. L., Wild, D. L. & Saldin, D. K. (2001). *Acta Cryst.* A**57**, 163–175.
Szöke, A. (1993). *Acta Cryst.* A**49**, 853–866.
Szöke, A., Szöke, H. & Somoza, J. R. (1997). *Acta Cryst.* A**53,** 291–313.
Terwilliger, T. C. (1999). *Acta Cryst.* D**55**, 1863–1871.
Terwilliger, T. C. (2000). *Acta Cryst.* D**56**, 965–972.
Terwilliger, T. C. & Berendzen, J. (1996). *Acta Cryst.* D**51**, 609–618.
Vellieux, F. M. D. A. P., Hunt, J. F., Roy, S. & Read, R. J. (1995). *J. Appl. Cryst.* **28**, 347–351.
Wang, B.-C. (1985). *Methods Enzymol.* **115**, 90–112.
Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.
Wilson, C. & Agard, D. A. (1993). *Acta Cryst.* A**49**, 97–104.
Xiang, S., Carter, C. W. Jr, Bricogne, G. & Gilmore, C. J. (1993). *Acta Cryst.* D**49**, 193–212.
Zhang, K. Y. J. (1993). *Acta Cryst.* D**49**, 213–222.
Zhang, K. Y. J., Cowtan, K. D. & Main, P. (1997). *Methods Enzymol.* **277**, 53–64.
Zhang, K. Y. J. & Main, P. (1990). *Acta Cryst.* A**46**, 41–46.